(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 August 2002 (22.08.2002)

PCT

(10) International Publication Number
WO 02/064617 A2

(51) International Patent Classification[7]: C07K

(21) International Application Number: PCT/US02/03486

(22) International Filing Date: 8 February 2002 (08.02.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0103295.2 9 February 2001 (09.02.2001) GB

(71) Applicant *(for all designated States except US)*: **ISIS INNOVATION LIMITED** [GB/GB]; British body coporate of Ewert House, Ewert Place, Summertown, Oxford OX2 7SG (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **STEPHENS, Matthew** [GB/US]; Department of Statistics, University of Washington, Box #354322, Seattle, WA 98195-4322 (US). **DONNELLY, Peter, James** [AU/GB]; Department of Statiscs, University of Oxford, 1 South Parks Road, Oxford OX1 3TG (GB). **SMITH, Nicholas, James** [GB/GB]; Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU (GB).

(74) Agent: **WILSON, Mary, J.**; Nixon & Vanderhye P.C., 1100 North Glebe Road, Suite 800, Arlington, VA 22201-4714 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/064617 A2

(54) Title: METHOD AND SYSTEM FOR HAPLOTYPE RECONSTRUCTION

(57) Abstract: A method for determining haplotype information from genotype information on individuals in a sample, comprises the steps of: executing a Markov chain Monte Carlo algorithm to derive information on the conditional distribution of haplotypes, based on the genotype information; and estimating haplotype information using the derived information on the conditional distribution. The method can also be used to quantify the uncertainty associated with the estimated haplotype information, using the derived information on the conditional distribution.

METHOD AND SYSTEM FOR HAPLOTYPE RECONSTRUCTION

The present invention relates to the field of haplotype reconstruction on the basis of genotype information.

Current routine genotyping methods typically do not provide haplotype information. Haplotype information is an essential ingredient in many analyses of fine-scale molecular genetics data, for example in disease mapping, or inferring population histories. Routine genotyping methods, such as DNA sequencing, typically do not provide phase information; that is both strands of a chromosome of a diploid individual are read simultaneously, but no information is obtained regarding on which strand a particular base resides. The choice of strand for each base or base sequence is known as the phase. The phase at each base (or base sequence) determines the haplotype. This phase information can be obtained, at considerable cost, experimentally, or (partially) through genotyping additional family members. Alternatively, a statistical method can be used to infer phase at linked loci from genotypes, and thus reconstruct haplotypes. The two most popular existing methods are maximum likelihood, implemented via the EM algorithm (Excoffier L, Slatkin M (1995), Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population, Molecular Biology and Evolution 12(5):921-927; Hawley M, Kidd K (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes, Journal of Heredity 86:409-411; Long et al. (1995) An EM algorithm and testing strategy for multiple locus haplotypes, American Journal of Human Genetics 56:799-810), and a parsimony method due to Clark (Clark A G (1990), Inference of haplotypes from PCR-amplified samples of diploid populations, Molecular Biology and Evolution 7(2):111-122).

Suppose one has a sample of n diploid individuals from a population. Let $G = (G_1,\ldots,G_n)$ denote the (known) genotypes for the individuals, $H = (H_1,\ldots,H_n)$ denote the (unknown) corresponding haplotype pairs, $F = (F_1,\ldots,F_M)$ denote the set of (unknown) population haplotype frequencies, and $f = (f_1,\ldots,f_M)$ denote the set of (unknown) sample haplotype frequencies (the M possible haplotypes are arbitrarily labelled $1,\ldots,M$).

The EM algorithm is a way of attempting to find the F that maximises the likelihood

$$L(F) = \Pr(G|F) = \prod_{i=1}^{n} \Pr(G_i|F) \quad (1)$$

Here

$$\Pr(G_i|F) = \sum_{(h_1,h_2)\in HAP_i} F_{h1}F_{h2}, \quad (2)$$

where HAP_i is the set of all (ordered) haplotype pairs consistent with the multilocus genotype G_i. Note that this likelihood assumes Hardy-Weinberg equilibrium (HWE). As a function of the population haplotype frequencies, this likelihood is just the probability (under HWE) of observing the sample genotypes.

The EM algorithm was implemented, for comparison with embodiments of the present invention (to be described below), to obtain an estimate $\hat{F}^{EM}$ for the population haplotype frequencies F, as described by Excoffier and Slatkin (1995). This was used as an estimate $\hat{f}^{EM}$ for the sample haplotype frequencies f (that is, $\hat{f}^{EM} = \hat{F}^{EM}$ was used). Since the estimate found by the EM algorithm typically depends on the starting point, for each data set the algorithm was applied using 100 different starting points, and took the estimate of F that gave the highest likelihood. Following Excoffier and Slatkin (1995), the first starting point was computed by finding all haplotypes that could occur in the sample given the genotypes, and setting each of these haplotypes to have equal frequency (it was found beneficial to add a small random pertubation to each frequency to avoid the algorithm converging to a saddle point in the likelihood). Each of the 99 other starting points was obtained by randomly sampling the frequencies of all possible haplotypes from a (multivariate) uniform distribution.

Although in theory the EM algorithm can be applied to any number of loci with any number of alleles, in practice implementations are limited by the need to store estimated haplotype frequencies for every possible haplotype in the sample. These storage requirements increase exponentially with the number of loci: for example, if any of the individuals is heterozygous at $\geq k$ loci then the number of possible haplotypes in the sample is $\geq 2^{k-1}$. In the implementation of the EM

algorithm, an arbitrary limit of 10^5 was imposed on the number of possible haplotypes, and the algorithm was not applied to data sets that exceeded this limit.

Within the maximum likelihood framework it is not clear how best to reconstruct the haplotypes themselves. The implementation of the EM algorithm used herein for the purposes of comparison with the embodiments of the invention, takes the approach of reconstructing haplotypes by choosing $\hat{H}^{EM}$ to maximise $\Pr(H|\hat{F}^{EM}, G)$, that is by choosing the most probable haplotype assignment given the genotype data and the estimated population haplotype frequencies $\hat{F}^{EM}$.

Clark's algorithm can be viewed as an attempt to minimise the total number of haplotypes observed in the sample, and hence as a sort of parsimony approach. The algorithm begins by listing all haplotypes which must, unambiguously, be present in the sample. This list comes from those individuals whose haplotypes are unambiguous from their genotypes: that is those individuals who are homozygous at every locus, or are heterozygous at only one locus. If no such individuals exist the algorithm cannot start (at least without extra information or manual intervention). Once this list of "known" haplotypes has been constructed, the haplotypes on this list are considered one at a time, to see if any of the unresolved genotypes can be resolved into a "known" haplotype plus a complementary haplotype. Such a genotype is considered resolved, and the complementary haplotype is added to the list of "known" haplotypes. The algorithm continues cycling through the list until all genotypes are resolved, or no further genotypes can be resolved in this way. The solution obtained can (and often does) depend on the order in which the genotypes are entered. In the comparisons the genotypes were entered once, in a random order, and cases were ignored where the algorithm could not start or completely resolve all genotypes.

When it successfully resolves all genotypes, Clark's algorithm results in an estimate, $\hat{H}^C$ of H. Sample haplotype frequencies f are estimated by the frequencies of the haplotypes reconstructed by the algorithm.

It is one object of the present invention to alleviate problems and shortcomings of previous algorithms.

Accordingly, the present invention provides a method for determining haplotype information from genotype information on individuals in a sample, comprising the steps of:

executing a Markov chain Monte Carlo algorithm to derive information on the conditional distribution of haplotypes, based on the genotype information; and

estimating haplotype information using the derived information on the conditional distribution.

The present invention utilises a new statistical method that improves on previous methods by exploiting ideas from population genetics and coalescent theory, which make predictions about the patterns of haplotypes to be expected in natural populations. A method according to the invention is Bayesian, allowing us to use these *a priori* expectations to inform haplotype reconstruction. It outperforms, and is more widely applicable than, existing algorithms: often error rates are reduced by >50% relative to its next-best alternative. A preferred feature is that it also estimates the uncertainty associated with each phase call and/or each entire haplotype reconstruction. This avoids inappropriate overconfidence in statistically-reconstructed haplotypes and, crucially, it allows subsequent experimental phase confirmation to be targeted effectively. Results suggest that in many cases the statistical method according to the invention is sufficiently accurate that reconstructing haplotypes experimentally, or by genotyping additional family members, may be an inefficient use of resources.

Another aspect of the present invention provides a system for determining haplotype information from genotype information on individuals in a sample, comprising:

an interface for receiving genotype information;

a module for executing a Markov chain Monte Carlo algorithm to derive information on the conditional distribution of haplotypes, based on the genotype information;

a module for estimating haplotype information using the derived information on the conditional distribution; and

an interface for outputting said haplotype information.

A further aspect of the present invention provides a computer program which is capable, when executed by a computer processor, of causing the computer processor to perform the above method.

The invention also provides a computer-readable storage medium having recorded thereon a computer program as defined above.

Embodiments of the invention will now be described, by way of example only, with reference to the accompanying drawings, in which:

Fig. 1 is a schematic illustration of the concept of reconstructing haplotypes to be like existing haplotypes;

Fig. 2 shows graphs comparing the accuracy of a method according to the invention with both the EM algorithm and Clark's method for short sequence data;

Fig. 3 shows graphs comparing the accuracy of a method according to the invention with the EM algorithm for microsatellite data; and

Fig. 4 is a graph comparing error rates of a method according to the invention with the EM algorithm for 100 simulated microsatellite data sets.

A haplotype reconstruction method according to the invention regards the unknown haplotypes as unobserved random quantities, and aims to evaluate their conditional distribution in light of the genotype data. To do this use is made of Gibbs sampling, a type of Markov chain Monte Carlo algorithm (MCMC algorithm, see Gilks WR, Richardson S, Spiegelhalter DJ (Eds.) (1996) Markov Chain Monte Carlo in Practice, London: Chapman & Hall, for further information), to obtain an approximate sample from the posterior distribution of H given G, $\Pr(H|G)$. Informally, the algorithm starts with an initial guess $H^{(0)}$ for H, and then repeatedly chooses an individual at random and estimates its haplotypes assuming all the other haplotypes are correctly reconstructed. Repeating this process sufficiently many times results in an approximate sample from $\Pr(H|G)$.

EMBODIMENT 1

Formally a method according to the invention involves constructing a Markov chain $H^{(0)}, H^{(1)}, H^{(2)}, \ldots$, with stationary distribution $\Pr(H|G)$, on the space of possible haplotype reconstructions, using the following algorithm.

Algorithm 1

Start with some initial haplotype reconstruction $H^{(0)}$. For $t = 0,1,2,\ldots$, obtain $H^{(t+1)}$ from $H^{(t)}$ using the following three steps:

1. Choose an individual i uniformly at random from all ambiguous individuals (i.e. individuals with more than one possible haplotype reconstruction.)

2. Sample $H_i^{(t+1)}$ from $\Pr(H_i|G,H_{-i}^{(t)})$, where H_{-i} is the set of haplotypes excluding individual i. $\Pr(H_i|G,H_{-i})$ is the conditional probability of the various possible haplotypes for individual i, given the genotype information for individual i and also assuming the haplotypes for all other individuals are known and are given by the list H_{-i}.

3. Set $H_j^{(t+1)} = H_j^{(t)}$ for $j = 1,\ldots,n, j \neq i$.

That this produces a Markov chain with the required stationary distribution follows from the proof for a general Gibbs sampler (see for example Gilks et al. 1996).

The difficulty in implementing the above algorithm lies in Step 2. Not only does the conditional distribution $\Pr(H_i|G,H_{-i})$, from which one is required to sample, depend on assumptions about the genetic and demographic models (or equivalently on a prior for the population haplotype frequencies F), but this distribution is not even known for most models (or priors) of interest. Nonetheless, it turns out to be helpful to rewrite the conditional distribution as follows. For any haplotype pair $H_i = (h_{i1}, h_{i2})$ consistent with genotypes G_i, one has

$$\Pr(H_i|G,H_{-i}) \propto \Pr(H_i|H_{-i}) \tag{3}$$

$$\propto \pi(h_{i1}|H_{-i})\pi(h_{i2}|H_{-i},h_{i1}) \tag{4}$$

where $\pi(\cdot|H)$ is the conditional distribution of a future-sampled haplotype given a set H of previously-sampled haplotypes. This conditional distribution is also not known in general. However, it is known in the particular case of *parent-independent mutation* (PIM), in which the type of a mutant offspring is independent of the type of the parent. Although this model is unrealistic for the kinds of system of interest (e.g. DNA sequence, multilocus microsatellite, and SNP data), it leads to a simple algorithm (see Embodiment 2) whose performance is roughly comparable to the EM algorithm (data not shown), and has at least two advantages over EM: it can be applied to very large numbers of loci, and it naturally captures the uncertainty associated with haplotype reconstructions. This simple algorithm also provides a convenient way of determining a good starting point for the improved algorithm that is now described.

The improved algorithm used by preferred embodiments of the present invention arises from making more realistic assumptions about the form of the conditional distribution $\pi(\cdot|H)$. Although for most mutation or demographic models, the conditional distribution $\pi(\cdot|H)$ is unknown, an approximation can be used. Formally, for a general mutation model with types in the countable set E, and (reversible) mutation matrix P, one such approximation, according to a particular embodiment, is

$$\pi(h|H) = \sum_{\alpha \in E} \sum_{s=0}^{\infty} \frac{r_\alpha}{r} \left(\frac{\theta}{r+\theta}\right)^s \frac{r}{r+\theta} \left(P^s\right)_{\alpha h} \qquad (5)$$

where r_α is the number of haplotypes of type α in the set H, r is the total number of haplotypes in H, and θ is a scaled mutation rate. Informally, this corresponds to the next sampled haplotype, h, being obtained by applying a random number, s, of mutations to a randomly chosen existing haplotype, α, where s is sampled from a geometric distribution. The approximation (5) arose from considering the genealogy relating randomly-sampled individuals, as described by the coalescent, and what it predicts about how similar a future-sampled chromosome is likely to be to those previously sampled. In particular, future-sampled chromosomes will tend to be more similar to previously-sampled chromosomes, as the sample size r increases, and as the mutation rate θ decreases.

The key to the increased accuracy of the algorithm is that the approximation (5) captures the idea that the next haplotype is likely to look either exactly the same as *or similar to* a haplotype that has already been observed; see Figure 1 for illustration. In this example, "similar to" means differing by one or a small number of mutational events. The particular embodiment of the method for haplotype reconstruction according to the invention, used herein for comparison purposes with previous methods, is based on substituting (5) into (4) to implement Step 2 of the Gibbs sampler. There are several other minor issues, both technical and practical (including for example how to estimate θ); details of these are given in Embodiment 3.

Figure 1 illustrates how a method according to the invention uses the fact that unresolved haplotypes will tend to be similar to known haplotypes. Suppose that one has a list of haplotypes, as shown in the figure, that are known without error (e.g. from family data, or because some individuals are homozygous). Then, intuitively the most likely pair of haplotypes for ambiguous individual 1 consists of two haplotypes that have high population frequency, as shown. All methods considered here will correctly identify this as the most likely reconstruction. However, ambiguous individual 2 cannot possess any of the haplotypes in the known list. The most plausible reconstruction for this individual consists of two haplotypes that are *similar, but not identical to* two haplotypes that have high population frequency, as shown. Of the methods considered here, only a method according to the invention uses this kind of information, leading to the improved performance observed.

For each run of the algorithm, R successive update steps were applied to obtain haplotype reconstructions $H^{(1)},\ldots,H^{(R)}$, the first b values of H were discarded as burn-in, and the remainder were thinned by storing the result every k iterations. For the simulation studies, where the method was applied to many data sets, relatively small values of R and b were used to keep the computational burden manageable: R=200,000, b=100,000, k=100. For the examples looked at here, much shorter runs produced similar average performance (data not shown), but for more complex problems larger values can be necessary to obtain reliable results. f is estimated by the mean of the empirical haplotype frequencies in the thinned sample, and methods outlined in Embodiment 3 are used to obtain a single point estimate,

$\hat{H}^{SSD}$, of H, together with estimates of $Q=(q_{ij})$, where q_{ij} denotes the probability that the phase call for individual i at locus j is correct.

EMBODIMENT 2

A basic version of the Gibbs sampler is considered here which arises from Algorithm 1 if one assumes PIM where the type of a mutant offspring is h with probability ν_h, independent of the type of the parent. In this case (for a constant-sized panmictic population) the conditional distribution $\pi(h|H)$ is known to be

$$\pi(h|H) = (r_h + \theta\nu_h)/(r + \theta) \qquad (7)$$

where r_h is the number of haplotypes of type h in H, r is the total number of haplotypes in H, and θ is the scaled mutation rate.

In principle one can substitute (7) into (4) to calculate (up to a normalising constant) $\Pr(H_i|G,H_{-i})$ for all possible values of H_i, and thus implement Step 2 of Algorithm 1. However this is impractical if the number of possible values of H_i is too large: if k denotes the number of loci at which individual i is heterozygous then there are 2^{k-1} different possible values for H_i, and if k is large the calculation becomes too complex. However, if one takes $\nu_h=1/M$ for all h, where M is the total number of different possible haplotypes that could be observed in the population, then one can solve these problems by exploiting the fact that, for those haplotype reconstructions H_i that do not contain any of the haplotypes in H_{-i}, the probabilities $\Pr(H_i|G,H_{-i})$ are all equal. This leads to Algorithm 2 below, which is practical for large samples and large numbers of loci.

Algorithm 2

Starting with an initial guess H for the haplotype reconstructions of all individuals, make a list consisting of the haplotypes $h=(h_1,\ldots,h_m)$ present in H, together with counts $r=(r_1,\ldots,r_m)$ of how many times each haplotype appears.

1. Pick an individual i uniformly at random, and remove its two current haplotypes from the list (h,r) (so the list now contains the haplotypes in H_{-i}). Let k be the number of loci at which i is heterozygous.

2. Calculate a vector $p=(p_1,...,p_m)$ as follows. For $j=1,...,m$ check whether the genotype G_i could be made up of the haplotype h_j plus a complementary haplotype, h' say. If not, set $p_j=0$, but if so search for h' in the list $(h_1,...,h_m)$. If h' is in the list, $h'=h_k$ say, then set $p_j=(r_j+\theta/M)(r_k+\theta/M)-(\theta/M)^2$, otherwise set $p_j=r_j(\theta/M)$.

3. With probability $2^k(\theta/M)^2/\left(\sum_j p_j + 2^k(\theta/M)^2\right)$ reconstruct the haplotype for individual i completely at random (i.e. by randomly choosing the phase at each heterozygous locus). Otherwise reconstruct the haplotype for individual i as h_j plus the corresponding complementary haplotype, with probability $p_j/\sum_{j'} p_{j'}$.

4. Add the reconstructed haplotype for individual i to the list (h,r).

The accuracy of this algorithm is similar to that of the EM algorithm. Note that the case $\theta v_h=1$ (for all h) corresponds to a uniform prior on the population allele frequencies F, and under this uniform prior the mode of the posterior distribution for F will be the same as the maximum likelihood estimate sought by the EM algorithm. This approach could thus be used to perform maximum likelihood estimation in problems that are too large for the EM algorithm.

EMBODIMENT 3

A more sophisticated Gibbs sampler is considered here which arises from using (5) and (4) to perform Step 2 of Algorithm 1. (In fact the algorithm presented is actually a "pseudo-Gibbs sampler", due to the fact that the conditional distributions sampled from are approximations that do not correspond to an explicit prior and likelihood.)

There are several problems to be overcome here. First, for multilocus data the expression (5) is not easy to compute, because the matrix P has the same dimension as the number of possible haplotypes, and is therefore potentially huge. Stephens M and Donnelly P (2000), Inference in molecular population genetics, Journal of The Royal Statistical Society, Series B 62:605-655, describe (in their Appendix 1) how to approximate (5) using Gaussian quadrature, and this approximation is made use of here. The approximation requires the specification of a mutation mechanism, and of a scaled mutation rate, θ_j, at each locus or site (note the contrast with θ in Stephens

M and Donnelly P (2000), which is the *overall* scaled mutation rate across sites or loci.

For the sequence data polymorphic sites were treated as linked biallelic loci, where a mutation at a locus causes the allele at that locus to change, and recurrent mutations are permitted. Non-polymorhic sites were ignored, because these sites add no ambiguity to the haplotypes (and, in any case the program used to simulate the sequence data outputs only the polymorphic sites.) For the method of this embodiment this is equivalent to setting $\theta_j = 0$ at non-polymorphic sites. For each polymorphic site in the sample the setting $\theta_j = 1/(\log(2n))$ was used, where n is the number of diploid individuals in the sample. This choice of θ_j gives, *a priori*, an expectation of approximately one mutation at each polymorphic site, during the ancestry of the sample since its most recent common ancestor. It also corresponds to an estimate, $\theta = S/(\log(2n))$, for the total scaled mutation rate across the region, where S is the number of polymorphic sites observed. This is, for moderate n, approximately Watterson's estimate for θ. To assess sensitivity of the results of choice of θ, other choices for θ_j at the polymorphic sites were looked at, in the range $\theta_j = 0.1$ to 1.0, for a few of the data sets with $n=50$. These values appeared to perform slightly less well, though the results were still a substantial improvement over the other known methods considered.

For the microsatellite data, a symmetric stepwise mutation model with 50 alleles and reflecting boundaries was used, the following setting of θ_j:

$$\theta_j = 0.5 \times \left(\left[1/(1 - H_j)^2 \right] - 1 \right),$$

where H_j is the observed heterozygosity at locus j.

Second, as in Embodiment 2 above, while in principle one can use (5) and (4) to calculate (up to a normalising constant) $\Pr(H_i | G, H_{-i})$ for all possible values of H_i, and thus implement Step 2 of Algorithm 1, instead the Gibbs sampler (Algorithm 1) is adjusted so that at each iteration one updates only a subset of the loci of a randomly chosen individual, as follows.

Algorithm 3

Start with some initial haplotype reconstruction $H^{(0)}$. For $t=0,1,2,\ldots$, obtain $H^{(t+1)}$ from $H^{(t)}$ using the following three steps:

1. Choose an individual i uniformly at random from all ambiguous individuals.

2. Select a subset S of ambiguous loci (or sites) in individual i to update. (In the absence of family, or other experimental data, the ambiguous loci are those for which individual i is heterozygous.) Let $H(S)$ denote the haplotype information for individual i at the loci in S, and $H(-S)$ denote the complement of $H(S)$, including haplotype information on all other individuals (so $H(S) \cup H(-S)=H$). Sample $H^{(t+1)}(S)$ from $\Pr(H(S)|G, H^{(t)}(-S))$.

3. Set $H^{(t+1)}(-S)=H^{(t)}(-S)$.

This modification of the algorithm does not affect its stationary distribution, regardless of how the subset S is chosen. In the specific example of the implementation used for this simulation study, S was formed by choosing 5 loci uniformly at random from the ambiguous loci in individual i (or all ambiguous loci if there were <5 ambiguous loci in individual i). At each iteration it is then necessary to compute $\Pr(H(S)|G, H^{(t)}(-S))$ for at most $2^5=32$ values of $H(S)$. (Note that, for $H(S)$ consistent with G,

$$\Pr(H(S)|G, H^{(t)}(-S)) \propto \Pr(H_i|H_{-i}),$$

and so the necessary probabilities can still be computed, up to a normalising constant, using (5).) Empirically it has been found that updating up to 5 loci at a time in this way produces reasonable mixing in small problems. However, the algorithm can have trouble mixing effectively for larger data sets, and choice of starting point can then be important. For the simulations a short preliminary run of Algorithm 2 was used (which updates all loci simultaneously in the chosen individual), from a random starting point, to provide a "good" starting point for Algorithm 3.

Finally, the information in the posterior distribution for H is summarised by a point estimate for H, and a matrix Q representing an estimate of the probability that each phase call is incorrect. This is done by specifying a loss function $Loss(\hat{H}^{SSD}, Q; H)$, which gives the loss for reporting the estimates $(\hat{H}^{SSD}, Q)$ when the true haplotypes are H, and attempting to minimise the posterior expected loss.

The particular loss function used in the exemplary implementation of the invention for this simulation study was

$$Loss(\hat{H}^{SSD}, Q; H) = -\max\left\{ \sum_{(i,j)\in C} \log(q_{ij}) + \sum_{(i,j)\notin C} \log(1-q_{ij}),\ \sum_{(i,j)\notin C} \log(q_{ij}) + \sum_{(i,j)\in C} \log(1-q_{ij}) \right\}$$

where $(i,j) \in C$ if the phase call in $\hat{H}^{SSD}$ for individual i at locus j is correct. There are other good summaries, corresponding to other sensible loss functions, for which the results are similar.

EMBODIMENT 4

As pointed out above, naïve implementation of Step 2 of Algorithm 1 requires computation of 2^{k-1} quantities, where k is the number of heterozygous sites in the individual being updated. For large values of k this is prohibitive, so to avoid this problem Embodiment 3 modified Step 2 to update only up to 5 sites at a time. This produces a valid algorithm that is tractable for large data sets. However, the present embodiment, Embodiment 4, provides an implementation of Step 2 that is computationally tractable for very large values of k, provided the number of sampled individuals is not too large. The computational cost of this new implementation increases only linearly with the total number of sites considered, and with the square of the total number of haplotypes in the sample. For smaller samples with very large numbers of sites this would be considerably more efficient than Embodiment 3.

The implementation of Step 2, according to the present embodiment, exploits the approximation to π (equation (5)), given in the appendix to Stephens M and Donnelly P (2000), Inference in molecular population genetics, Journal of The Royal Statistical Society, Series B 62:605-655, using Gaussian quadrature with Q quadrature points $(T_1,\ldots,T_Q)$ and associated quadrature weights $(W_1,\ldots,W_Q)$. This approximation allows $\Pr(H_i|G,H_{-i})$ to be written in the form of a product over sites, summed over latent variables $(n_1,n_2,c_1,c_2,t_1,t_2)$:

$$\Pr(H_i|G,H_{-i})=\sum_{(n_1,c_1)}\sum_{(n_2,c_2)}\sum_{t_1=1}^{Q}\sum_{t_2=1}^{Q}\Big[\Pr(n_1,c_1,n_2,c_2,t_1,t_2|G,H_{-i})$$

$$\prod_{r=1}^{R}\Pr\big((h_{i0r},h_{i1r})|G,H_{-i},n_1,c_1,n_2,c_2,t_1,t_2\big)\Big]$$

$$=\sum_{(n_1,c_1)}\sum_{(n_2,c_2)}\sum_{t_1}\sum_{t_2}\frac{1}{(2N-2)}\frac{1}{(2N-1)}W_{t_1}W_{t_2}$$

$$\prod_{r=1}^{R}F\big(h_{n_0,c_0,r},h_{i,0,r};\theta,T_s,2N-2\big)F\big(h_{n_1,c_1,r},h_{i,1,r};\theta,T_t,2N-1\big) \tag{8}$$

where R is the total number of loci/sites, $h_{n,c,r}$ is the allele carried by individual n at locus r on haplotype c ($c = 0, 1$), the sum over indices (n_1,c_1) is over all haplotypes in H_{-i}, the sum over indices (n_0,c_0) is over all haplotypes in $H_{-i} \cup h_{i,0}$, N is the effective population size and θ is the scaled mutation rate, and

$$F(i,j;\theta,t,n)=\sum_{m=0}^{\infty}\frac{(\theta t/n)^m}{m!}\exp(-\theta t/n)\big(P^m\big)_{ij} \tag{9}$$

Both the accuracy and the computational requirements of the approximation increase with the number of quadrature points Q used. The currently preferred default value is $Q = 4$, but using $Q = 1$ or $Q = 2$ would speed computation. The sum (9) is approximated by its first 51 terms.

Using the representation (8), sampling from $\Pr(H_i|G,H_{-i})$, as required for step 2, can be achieved by first sampling the latent variables (n_1,n_2,t_1,t_2) from $\Pr(n_1, n_2, t_1, t_2|G, H_{-i})$ (which requires $o(n^2Q^2R)$ computation, where n is the total number of chromosomes in the sample), and then sampling (h_{i0r},h_{i1r}) from $\Pr((h_{i0r},h_{i1r})|n_1,n_2,t_1,t_2)$ independently for each r.

Features of the invention described with reference to one embodiment can, of course, be used as appropriate with other embodiments. For example, one preferred

implementation of the invention is to use algorithm 3, but deriving the conditional distribution at the end of Step 2 from the formula (5).

Two further computational techniques are preferably employed to improve computational performance of embodiments of the invention. The first is to store in an array the number of differences between every pair of haplotypes at biallelic loci. Only a small proportion of the elements of this array will change during each iteration; these are updated at the end of each iteration. The second is to pre-compute values for the function F described above, and for powers of F, and store them in a look-up table on program initiation. Use of either or both of these two techniques substantially reduces the computational time required to complete either the new Step 2 (as in Embodiment 4) or the previous Step 2.

Methods embodying the invention can be implemented as computer software comprising computer instructions or code, stored on a computer-readable storage medium. The invention can be embodied as a system for executing such software, using a computer processor. The system may include an interface for receiving genotype information, for example from a user interface, or from a storage medium, or over a communications network, such as the Internet. The system may also comprise modules for performing a method embodying the invention. The system may include an interface for outputting resulting haplotype information, for example by displaying it on a screen, printing it, storing it on a storage medium or sending it over a communications network. Methods embodying the invention may, of course, have the initial step of obtaining the genotype information, either by using known experimental techniques, or by inputting genotype data previously obtained, for example by downloading such genotype data over a communications network.

RESULTS

To compare the performance of the statistical methods of haplotype reconstruction, various types of DNA sequence, and tightly linked multilocus microsatellite, data with known phase, were simulated. (The details of how these data were simulated are explained with reference to the relevant Figures.) Simulated

haplotypes were randomly paired, and the methods were compared on their ability to reconstruct these known haplotypes from the resulting genotype data, in which phase information is ignored. There are many possible aims for statistical methods of haplotype reconstruction. Two particular tasks were concentrated on here:

(I) reconstructing the haplotypes of sampled individuals, which is the main focus of Clark (1990);

(II) estimating sample haplotype frequencies, which is the main focus of Excoffier and Slatkin (1995)

For (I) performance was measured by the *error rate*, being the proportion of individuals with ambiguous phase whose haplotypes are incorrectly inferred. For (II) the *discrepancy* was used between the estimated and true sample haplotype frequencies:

$$D(\hat{f};f)=\frac{1}{2}\sum_j|\hat{f}_j-f_j|, \qquad (6)$$

with summation over all possible haplotypes, where $\hat{f}_j$ and f_j denote respectively the estimated and true sample frequency of the jth haplotype. The discrepancy is equivalent to the I_f score used by Excoffier and Slatkin (1995): $D=1-I_f$.

The results of comparisons, shown in Table 1 and Figures 2-3, demonstrate that the accuracy of the method according to the invention substantially improves on both the EM algorithm and Clark's method, with mean error rates often reduced by >50%. Not only is the average performance improved, but this improvement is achieved by a consistent improvement across many data sets, rather than an extreme improvement in a minority of cases. For example, Figure 4 shows that for the simulated microsatellite data sets with n=50 and R=4, the EM algorithm gave a smaller error rate than a preferred method embodying the invention for only 3 data sets out of 100.

Table 1

	Mean error rate	(standard error)
Clark's method	0.42	(0.03)
Method of the invention	0.20	(0.02)

Table 1 shows comparison of the accuracy of a preferred method embodying the invention vs Clark's method for long sequence data (~60-100 segregating sites (s.s.); e.g. recent studies report 78 s.s. in a 9.7kb region and 88 s.s. in a 24-kb region). Results are averages over 15 simulated data sets, each of n=50 individuals, simulated with $\theta=4N_e\mu=16$, and $R=4N_er=16$, where N_e is the effective population size, μ is the total per generation mutation rate across the region sequenced, and r is the length, in Morgans, of the region sequenced. (20 independent data sets were simulated for a constant-sized panmictic population under the infinite sites model, with recombination, using a coalescent-based program. Each simulated data set consisted of 100 haplotypes randomly paired to form 50 genotypes. Clark's algorithm produced a unique haplotype reconstruction for 15 of these, and the other 5 were discarded.) Implementations of the EM algorithm cannot typically cope with these kinds of data, as the number of possible haplotypes is too large.

Figure 2 shows a comparison of accuracy of a preferred method embodying the invention (solid line) vs EM (dotted line) and Clark's method (dashed line) for short sequence data (~5-30 segregating sites). Top row: mean error rate for haplotype reconstruction; bottom row: mean discrepancy for estimation of haplotype frequencies. Data sets of $2n$ haplotypes were simulated, randomly paired to form n genotypes, under an infinite sites model, with θ=4, and different assumptions about the local recombination rate R (R and θ are defined in the explanation of Table 1), using a coalescent-based program. For each combination of parameters considered, 100 independent data sets were generated, and those data sets for which the total number of possible haplotypes was $>10^5$ (the limit of the implementation of the EM algorithm) were discarded, which typically left >90 data sets on which to compare the methods. Each point thus represents an average over around 90-100 simulated data sets. Horizontal lines above and below each point show approximate 95% confidence intervals for this average (±2 standard errors). The results for Clark's algorithm for R=40 are omitted because there was difficulty getting the algorithm consistently to provide a unique haplotype reconstruction for these data.

Figure 3 shows a comparison of accuracy of a preferred method embodying the invention (solid line) vs EM (dotted line) for microsatellite data. Top row: mean error rate for haplotype reconstruction; bottom row: mean discrepancy for estimation of haplotype frequencies. Data sets of $2n$ haplotypes were simulated, randomly paired to form n genotypes, for 10 equally-spaced linked microsatellite loci, from a constant-sized population, under a symmetric stepwise mutation model, using a coalescent-based program. It was assumed that $\theta=4N_e\mu=8$ (where μ is the per generation mutation rate per locus, assumed constant across loci) and various values for the scaled recombination rate between neighbouring loci, $R=4N_er$, where N_e is the effective population size, and r is the genetic distance, in Morgans, between loci. (For example, for humans, assuming $N_e=10^4$, and the genome-wide average recombination rate, 1cM=1Mb, the right-hand column would correspond to 10kb between loci.) For each combination of parameters considered, 100 independent data sets were generated. Each point thus represents an average over 100 simulated data sets. Horizontal lines above and below each point show approximate 95% confidence intervals for this average (±2 standard errors). There was difficulty getting Clark's algorithm consistently to provide a unique haplotype reconstruction for these data.

Figure 4 is a graph of error rates using a preferred method embodying the invention (solid line) and the EM algorithm (dotted line) for each of the 100 simulated microsatellite data sets with $n=50$ and $R=4$. The EM algorithm gives a smaller error rate than a preferred method embodying the invention for only 3 of the 100 data sets.

An important feature of a preferred method embodying the invention is that it quantifies the uncertainty in its phase calls at each ambiguous site by outputting an estimate of the probability that each call is correct. Table 2 shows the method to be well calibrated, in that, on average, phases called with x% certainty are correct approximately x% of the time. This is another substantial advantage of a preferred method embodying the invention, and in this sense the performance improvements illustrated in Figures 2-3 and Table 1 under-represent the gains achieved.

Table 2 gives the results of calibration tests. Table entries show, for the simulated data sets used for Table 1 and Figures 2-3, the proportion of all phase calls

at ambiguous loci or sites, made with a given degree of confidence, that were actually correct. (The phase call in individual i at locus j was regarded as incorrect if the alternative call would give strictly fewer differences between the true and estimated haplotypes; otherwise the call was regarded as correct. Formally the entries in each row of the table show $\#\{(i,j): x\% < q_{ij} \le (x+10)\% \cap$ Phase call for individual i at locus j is correct$\} / \#\{(i,j): x\% < q_{ij} \le (x+10)\%\}$, for x=50,60,70,80,90, where $\#A$ denotes the number of members of the set A.) The results suggest that a method according to the invention tends, on average, to be slightly conservative in its estimate of the probability of having made a correct call.

Table 2

		Estimated probability of correct call				
		0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Long sequence data		0.59	0.82	0.82	0.82	0.99
Short sequence data	R=0	0.58	0.82	0.87	0.89	0.95
	R=4	0.60	0.86	0.88	0.90	0.98
	R=40	0.62	0.72	0.77	0.84	0.96
Microsatellite data	R=0	0.60	0.73	0.81	0.87	0.99
	R=2	0.60	0.69	0.78	0.86	0.98
	R=4	0.60	0.70	0.76	0.83	0.97

The effect on a method according to the invention of departures from Hardy-Weinberg equilibrium (HWE) in data was assessed. Published haplotype data from a geographically dispersed/distinct sample were randomly combined in pairs, first under HWE and then in a way that respected geographical structure. Table 3 shows that these departures from HWE have little effect. The type of geographical structure modelled will tend to increase the amount of homozygosity in the sample, which tends to reduce the number of ambiguous individuals. Deviations from HWE in the other direction (an increased proportion of heterozygotes) will tend to make haplotype reconstruction more difficult for all methods.

Table 3

	Mean error rate	Mean discrepancy
Data set HW	0.21 (0.006)	0.16 (0.002)
Data set NHW	0.21 (0.008)	0.16 (0.004)

Table 3 illustrates the effect on a preferred method embodying the invention of deviations from Hardy-Weinberg equilibrium in data. Results are averages over 20 simulated data sets, with standard errors for this average in parentheses. Data set HW was formed under Hardy-Weinberg equilibrium, and NHW formed under an assumption of geographical structure (see below). The results suggest that deviations from Hardy-Weinberg equilibrium have little effect on the average performance of a method according to the invention (though the performance is slightly more variable). To create the data sets HW and NHW use was made of published, experimentally-determined haplotype data from a 3kb region of the beta-globin gene, sequenced in 253 chromosomes (Harding et al. (1997), Archaic African and Asian Lineages in the Genetic Ancestry of Modern Humans, American Journal of Human Genetics 60:772-789, data from their appendix A). Six subpopulations were represented in the sample: Vanuatu, Papua New Guinea, Sumatra, the Gambia, the UK, and the Nuu-Chah-Nulth. These data were used to create HW and NHW by repeating the following procedure 20 times: a) remove one chromosome at random from each subpopulation with an odd number of chromosomes in the sample (leaving 250 haplotypes); b) form 125 genotypes by i) (for HW) randomly pairing all haplotypes, and ii) (for NHW) randomly pairing haplotypes within each subpopulation.

Haplotypes are the raw material of many genetic analyses, but the rapid growth in high-throughput genotyping techniques has not been matched by similar advances in cheap experimental haplotype determination. The present invention introduces a new statistical method for haplotype reconstruction that has three major

advantages over existing statistical methods: increased accuracy, wider applicability, and the facility to assess accurately the uncertainty associated with each phase call.

The key to the increased accuracy is the use, in addition to the likelihood, of the fact that, *a priori*, unresolved haplotypes tend to be similar to known haplotypes (see Figure 1). The particular quantitative way in which one captures this prior expectation is motivated by coalescent theory. It amounts to specifying a statistical model (or, depending on your philosophy, a "prior") for the population genetics aspect of the problem, namely the results of the evolutionary process that generated the haplotypes in the first place. Of course one would expect the method to perform well if this is exactly the model which is generating the data, but this will never be the case for real data. Thus what matters in practice is whether the model used, and the implicit prior it induces on haplotype structure, does a reasonable job of capturing important features of the haplotype structure in real data. If so, then one would expect the method to perform well, and to outperform methods (including those to which it is compared here) which do not model the haplotype structure in the population.

Unfortunately there simply do not exist enough real data sets, with known haplotypes for sequence or closely linked markers, to allow sensible statistical comparisons of different methods. However, coalescent methods have proved useful for a wide range of molecular genetics data, and so it seems reasonable, *a priori*, to expect their use here to helpfully capture some of the key population genetics aspects of real data. Further, the simulation results provide evidence that the performance of a method according to the invention is relatively robust to deviations in data from the underlying modelling assumptions. It is noted in particular that: i) the data underlying Table 1 and Figure 1 were simulated under a mutation model different from that used to derive the prior in a method according to the invention; ii) the majority of data sets were simulated with recombination, which is not included in the model; and iii) in none of the tests of a method according to the invention did one use the actual parameter values under which data were generated (as described in Embodiment 3, these were estimated within the method via simple summary statistics). Despite these deviations from the model, the method performed well. Thus, although the method makes more explict assumptions than the other methods

considered, it would be a mistake to conclude that it *requires* all these assumptions hold for it to provide a useful improvement in performance. (It is however true that in analysing some real microsatellite data according to the method of the invention, it can be prudent to drop the assumption of a strict stepwise mutation mechanism.)

In addition to the modelling assumptions underlying a preferred method embodying the invention, the likelihood used here, and by the EM algorithm, assumes Hardy-Weinberg equilibrium (HWE). Fallin D and Schork NJ (2000), (Accuracy of Haplotype Frequency Estimation for Biallelic Loci, via the Exception-Maximization Algorithm for Unphased Diploid Genotype Data, American Journal of Human Genetics 67:947-959) provide a good discussion of the general consequences of departures from HWE in data, and show that the EM algorithm can still give good results when HWE is not strictly satisfied. The results in Table 2 suggest that geographical structure of the sort plausible for human populations does not affect the average accuracy of a method according to the invention. In the light of this it would be a misunderstanding to assume that either method "relies" on HWE for its validity; both methods are better thought of as "black box" estimation methods, and it is appropriate to assess their performance for data generated under a range of scenarios, as done here.

A method according to the invention is also directly applicable to SNP data. In fact, the final column of Figure 2 corresponds to SNP data (10-30 SNPs over 100kb, assuming 1cM=1Mb) in which ascertainment of SNPs is independent of their sample frequencies. Although different studies use different ascertainment strategies, most of these strategies tend to preferentially select SNPs with higher frequencies. Higher-frequency variants produce, on average, a larger total number of ambiguous phases (through greater heterozygosity), but these phases are typically easier to estimate statistically than those of lower frequency variants (for example, there is no information about phase for variants that appear only once in the sample). For these reasons one would expect that when applied to real SNP data, although a method according to the invention would typically give more incorrect phase calls (in absolute terms) than for the corresponding SNP data simulated without taking ascertainment into account, a higher *proportion* of ambiguous phases would be called correctly.

As well as being more accurate, a method according to the invention is also more widely applicable than other available methods. Existing implementations of the EM algorithm are limited in the size of problem they can tackle. For example, they are typically impracticable for sequence data containing individuals whose phase is ambiguous at more than ~30 sites. Similarly, they cannot cope with large numbers of linked SNPs. Clark's algorithm can deal with very long sequences (or large numbers of SNPs), but may fail either to start or to resolve all genotypes completely. These problems with Clark's algorithm arose in many of the settings examined. In contrast, a method according to the invention suffers from neither limitation, although the running time required will increase with the size and complexity of the problem. For the simulated data sets considered, the running time for our method ranged from a few minutes to a few hours (an a PC with a 500 MHz processor), while the EM algorithm and Clark's method typically took only a matter of seconds for those problems to which they could successfully be applied. However, since a few hours of calculation on a computer is small relative to the costs of data collection and experimental haplotype reconstruction, this kind of difference in speed is not a particularly important consideration in this context. Given reasonable modern computing resources, the method of the invention is practicable for at least hundreds of individuals genotyped at 100 sites or more.

An important and further novel feature of a preferred method embodying the invention is that it provides estimates of the uncertainty associated with each phase call. Quantifying uncertainty is, of course, good practice in any statistical estimation procedure. Formally, a method according to the invention provides a sample from a distribution over possible haplotype reconstructions. While it might be tempting to hope that one could summarise the posterior distribution by a few most common configurations, in the example cases studied the support is typically spread rather thinly over an enormous number of possible haplotype configurations. Therefore the choice was made to summarise the posterior distribution by a single "best" phase call at each position, and an estimate of the marginal probability that each phase call is correct. In practice this risks discarding some of the information in the posterior distribution, particularly complex dependencies between the phase calls at different positions both within and between individuals. Nonetheless, this summary is a

helpful way of visualising the full joint distribution over possible haplotype reconstructions.

Unless haplotype reconstruction is an end in itself, it is natural to make use of a sample from the posterior distribution of haplotype reconstructions in subsequent analyses. Any statistical procedure that uses haplotype data can easily be applied independently to several sampled haplotype reconstructions. For certain inference problems, particularly when using Bayesian methods, which provide posterior distributions over parameters of interest, uncertainty in the haplotype reconstruction can then be taken into account by averaging results of the independent analyses. However, in many inference problems (e.g. estimating recombination rates) it would be preferable to develop a method for jointly inferring haplotypes and parameters of interest, and in other settings (e.g. significance testing) the best way to combine the results of independent analyses is far from clear. As a practical general solution one option is performing independent analyses using 10 sampled haplotype reconstructions, to investigate the robustness of conclusions to inferred haplotypes. If conclusions differ among the 10 analyses, then experimental methods for haplotype reconstruction may be required to confirm findings.

Statistical methods can be used in conjunction with experimental methods to provide more accurate estimates of individual haplotypes. Although Clark's algorithm has been treated as an automatic method for haplotype reconstruction, it has often been used as an exploratory tool to suggest putative haplotype reconstructions, which could then be confirmed by allele-specific PCR. This seems a powerful approach, and a preferred method embodying the invention can also be used in this way. The ability of a preferred method embodying the invention to accurately assess the uncertainty associated with the phase call at each individual site (or locus) gives it the substantial practical advantage of allowing experimental effort to be directed at sites or loci whose phases are most difficult to reconstruct statistically. Software embodying the invention can use experimentally-verified phase information (for either complete individuals, or specific sites or loci) in estimating the unknown haplotypes, and this will usually produce a further substantial reduction in error rate.

The availability of family data can also improve statistical estimation of haplotypes. In the case where triples of mother, father and child have been collected, the child's genotype information can be used to infer the parental haplotypes at many loci or sites, and (provided the genetic distance across the region is not too large) a preferred method embodying the invention can then use this known phase information in estimating the remaining ambiguous phases. For data from extended pedigrees the situation is often more complex. Current methods for haplotype reconstruction in pedigrees ignore population genetic considerations, and rely on information in transmission events to infer haplotypic phase. These tranmission events carry information on haplotypes over much larger genetic distances than considered here, and pedigree methods can thus be effective even for widely-spaced markers. However, the use of ideas from population genetics to model founder haplotypes in such pedigrees could lead to worthwhile performance improvements, particularly as denser maps of markers become available.

Another simulation study has assessed the performance of the EM algorithm for reconstructing phase from genotype data at linked biallelic loci, and considered 5 linked biallelic loci. The general conclusion was that the performance of the EM algorithm is good. However examples using the present invention have considered rather "larger" problems. The following different data have been considered: long sequence data (60-100 segregating sites, Table 1); short sequence data (5-30 segregating sites, Figure 2); and 10 linked microsatellite loci (Figure 3). Haplotype reconstruction is much harder for these larger (but realistic) problems, and the improvement made by a method according to the present invention over EM is practically important.

A more technical, less substantial, methodological difference between the approaches is that the examples of the present invention assessed accuracy of haplotype sample frequency estimates using the discrepancy, while other studies use the mean squared error (MSE). The MSE has the drawback that its range of possible values is strongly dependent on the number of possible haplotypes, making comparisons among data sets with differing numbers of loci harder to interpret. A similar measure, which does not have this drawback, is the sum of squared errors (SSE). Using MSE to measure accuracy of haplotype sample frequency estimates

gives results which seem to favour a method according to the invention more strongly than using the discrepancy.

A preferred method embodying the invention does not use information about genetic distances between loci or sites, and is best suited to cases where the loci are tightly linked. Nonetheless, our simulation results show that it continues to perform well in the presence of moderate amounts of recombination (loci or segregating sites spread over about 100kb in humans, provided there are no recombination hot-spots in the region). A preferred method embodying the invention can be extended to deal with loci or sites spread over larger genetic distances by replacing the approximation (5) or (8) with an approximation that takes genetic distance between loci into account.

A common and important problem with MCMC algorithms is knowing how long one needs to run the algorithm to obtain reliable results (often this problem is referred to in terms of diagnosing "convergence" of the Markov chain.) Checking (and indeed, attaining) convergence is in general notoriously difficult, so it is important to note that the method of the present invention does not necessarily need to "converge" in order to improve on the accuracy of other methods. In particular, the results of simulation experiments using the present invention did not rely on checking convergence of the algorithm, and so provide direct evidence that, for the size of problem considered, the runs were sufficiently long to give a substantial average gain in accuracy over other methods, regardless of whether the Markov chain had actually "converged" in each instance. Nonetheless it is helpful to be aware of potential problems caused by lack of convergence. The main danger is that the algorithm gets "stuck" in a local mode of the posterior distribution of haplotype reconstructions, and fails to find other, perhaps more strongly supported, modes. Depending on the severity of the problem, this kind of behaviour can be difficult to identify on the basis of a single run of the algorithm, since this run could remain stuck in a single local mode for a very long time, and give no clue that other modes exist. Therefore a preferred aspect of embodiments of the invention is to investigate convergence using multiple runs of the algorithm with different initial values for the seed of the random-number generator. If the algorithm tends to get stuck in local modes then these different runs may give qualitatively quite different results,

effectively diagnosing the problem. (Although more formal approaches are possible, most are based on diagnosing convergence for a small set of continuous parameters, which is not the situation here). Where possible the length of the runs should be increased until they all give qualitatively similar results. If this proves impractical then further experimental investigation may be necessary to decide between the solutions.

As mentioned above, one problem with the previous techniques is that if multiple runs give different results, then it is difficult to know which of the results is most likely to be accurate. A preferred feature, for use with any embodiment of the invention, is to introduce a measure of the "goodness-of-fit" of estimated haplotypes to a coalescent model, to help alleviate this problem. The goodness-of-fit measure preferably used is the pseudo-likelihood (see Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, series B 36,* 192-236) of H:

$$\prod_{i=1}^{n} \Pr(H_i | H_{-i}),$$

where the conditional probability $\Pr(H_i | H_{-i})$ is computed using the same approximation as is used in the main algorithm of the embodiment of the invention.

A preferred method embodying the invention can usefully be extended to allow it to deal with missing genotype data in some individuals at some loci. This is straightforward in principle, by augmenting the space that the MCMC scheme explores to include the missing data. Similarly the method can be extended to allow for the possibility of genotyping error. For realistic amounts of missing data, and realistic genotyping error probabilities, these extensions are unlikely to greatly increase the computational time a method according to the invention requires. (Indeed incorporating the possibility of genotyping error may even provide a way to improve the mixing of the MCMC scheme, as it will tend to flatten out modes in the posterior distribution.)

One specific way to deal with missing genotype data, for use with the above embodiments of the invention, is to augment the space that the MCMC scheme

explores to include the missing genotypes. To describe the details of this implementation we introduce a slight change in notation. Let $G = (G_1,...,G_n)$ denote the *observed* genotype data, $\overline{G} = (\overline{G}_1,...,\overline{G}_n)$ denote the *unobserved* (missing) genotype data, $H = (H_1,...,H_n)$ denote the haplotype data corresponding to the observed genotype data, and $\overline{H} = (\overline{H}_1,...,\overline{H}_n)$ denote the haplotype data corresponding to the unobserved genotype data. Thus $(H,\overline{H})$ is the complete haplotype data at all sites in all individuals. The new algorithm simulates a Markov Chain $(H^{(0)},\overline{H}^{(0)})$, $(H^{(1)},\overline{H}^{(1)})$,... with stationary distribution Pr($H, \overline{H} \mid G$). This is achieved by using the full genotype data $(G,\overline{G})$ in steps 1 and 2 of the algorithm, and by replacing Steps 1 to 3 of the algorithm (of any preceding embodiment) with the following step for at least some of the iterations:

Step 4: Choose an individual i at uniformly random from all individuals. Let $\overline{H}_{(i,c)}$ denote the unknown ("missing") haplotype data in individual i, chromosome c. Sample $\overline{H}_{(i,0)}^{(t+1)}$ from $\Pr\left(\overline{H}_{(i,0)} \middle| H^{(t)}, G^{(t)}, \overline{H}_{-i}^{(t)}\right)$; sample $\overline{H}_{(i,1)}^{(t+1)}$ from $\Pr\left(\overline{H}_{(i,1)} \middle| H^{(t)}, G^{(t)}, \overline{H}_{-i}^{(t)}, \overline{H}_{(i,0)}^{(t+1)}\right)$; and set $\overline{H}_{-i}^{(t+1)}$ to $\overline{H}_{-i}^{(t)}$. Update $\overline{G}^{(t+1)}$ to be consistent with $\overline{H}^{(t+1)}$ ($\overline{H}^{(t+1)}$ completely determines $\overline{G}^{(t+1)}$).

The two sampling steps required to execute Step 4 are performed using the analogue of expression (8) for just one chromosome (in the case of Embodiment 4, or equivalent equations for other embodiments). (For just one chromosome, this computation scales only linearly in n and in the number of sites at which genotype data are missing.)

The preferred method for handling missing (unknown) genotype data is to alternate between performing steps 1 to 3, and performing step 4, such that on each iteration either steps 1 to 3 are performed or step 4 is performed. Of course, it is also possible to perform step 4 more or less often than every alternate iteration, for example to achieve optimal processing.

An alternative approach to the whole problem would be to assume more explicit models for mutation, recombination, and population demography, and to jointly estimate haplotypes with the parameters of these models. Kuhner MK and

Felsenstein J (2000) (Sampling among haplotype resolutions in a coalescent-based genealogy sampler, Genetic Epidemiology 19 (Suppl 1):S15-S21) describe an MCMC scheme which could be used to study the conditional distribution of haplotypes given genotypes, but they do not disclose providing estimates of haplotype information. Their MCMC scheme makes explicit use of the genetic distance between markers, and is a modified version of the scheme for known haplotypes described in Kuhner MK, Yamato J and Felsenstein J (2000), Maximum likelihood estimation of recombination rates from population data, Genetics 156:1393-1401. However, because of its computational complexity this scheme is currently practical only for very small genetic distances, and so could not usefully be applied to most of the data sets considered here.

In conclusion, for many of the data sets considered here, the statistical method according to the invention succeeded in correctly reconstructing the haplotypes of at least 80% of the sample. Although explicit conclusions will depend on power calculations for specific analyses, the accuracy of a method according to the invention suggests that in many settings the optimal use of experimental resources will be to maximize the number of unrelated individuals genotyped. In others, it will be most efficient to target experimental effort on those phase calls that are identified as having a moderate probability of being wrong, or that are critical to the conclusions of the study.

Regarding the utility and industrial applicability of the present invention, as already mentioned, the haplotype information obtained by the method can be used for example in disease mapping or inferring population histories, or for optimizing the allocation of experimental resources in genotyping most effectively, and so on. Haplotype information can be useful in interpretting clinical trials data, for example it can be useful to know if possession of certain haplotypes makes people more or less likely to respond to a drug. Haplotype information can also be used in prenatal diagnosis or prediction of the likelihood of inheriting a disease. Some specific examples of applications are as follows:

Population histories have been inferred from haplotype information (Harding et al. (1997), Archaic African and Asian Lineages in the Genetic Ancestry of Modern Humans, American Journal of Human Genetics 60:772-789);

Statistically-estimated haplotypes have been used to help identify a gene involved in type-2 diabetes (*Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus*, Horikawa Y, Oda N, Cox NJ et al (2000), Nature Genetics 26:163-175);

Haplotypes (estimated from family data) have been used to identify a genetic variant associated with adult-type hypolactasia (Enattah N S et al (2002), Nature Genetics 30:233-237), and similarly for Crohn's disease (*Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease*, Rioux J D et al (2001), Nature Genetics 29:223-228); and

Statistically-estimated haplotypes have also been used to examine the association between the IL8 gene and bronchiolitis (*Unusual Haplotypic Structure of IL8, a Susceptibility Locus for a Common Respiratory Virus*, Hull J, Ackerman H, Isles K, et al, AM J HUM GENET 69(2):413-419 (Aug 2001)).

CLAIMS

1. A method for determining haplotype information from genotype information on individuals in a sample, comprising the steps of:

executing a Markov chain Monte Carlo algorithm to derive information on the conditional distribution of haplotypes, based on the genotype information; and

estimating haplotype information using the derived information on the conditional distribution.

2. A method according to claim 1, further comprising the step of quantifying the uncertainty associated with the estimated haplotype information using the derived information on the conditional distribution.

3. A method according to claim 2, wherein said quantifying step includes quantifying the uncertainty associated with each phase call in the haplotype estimation.

4. A method according to claim 2 or 3, wherein a loss function is used for evaluation in the quantifying step.

5. A method according to any one of the preceding claims, wherein a loss function is used for evaluation in the estimating step.

6. A method according to any one of the preceding claims, wherein said executing step is performed using a Gibbs sampler.

7. A method according to any one of the preceding claims, further comprising the step of specifying a prior distribution on haplotypes.

8. A method according to any one of the preceding claims, further comprising the step of specifying the conditional distribution of the haplotypes in an individual

for which genotype information is not available, given known haplotype information in a sample of individuals from a population.

9. A method according to any one of the preceding claims, comprising the steps of: choosing an individual from a sample for which genotype information has been obtained; sampling the haplotypes for assigning to that individual from the conditional distribution of haplotypes given the genotype information for that individual and assuming the haplotypes of all other individuals have been correctly reconstructed.

10. A method according to claim 9, further comprising reassigning the whole haplotype for the chosen individual, on the basis of the haplotypes sampled from the conditional distribution.

11. A method according to claim 9, further comprising reassigning the phase only at predetermined loci or sites of the haplotypes of the chosen individual, on the basis of the haplotypes sampled from the conditional distribution.

12. A method according to claim 9, 10 or 11, wherein the step of sampling the haplotypes is based on the haplotype frequencies from all other individuals in the sample, excluding the chosen individual.

13. A method according to any one of claims 9 to 12, comprising iterating the choosing and estimating steps.

14. A method according to claim 13, comprising the steps of:

repeatedly: storing the haplotype assignment information after a predetermined number of iterations then continuing iterating; thereby accumulating a stored set of different possible haplotype assignments;

using the set of stored haplotype assignments to obtain an estimate of the uncertainty in the assigned haplotype information.

15. A method according to any one of the preceding claims, further comprising the step of outputting haplotype information for an individual in the sample.

16. A method according to any one of the preceding claims, further comprising the step of outputting information on all haplotypes in the sample.

17. A method according to any one of the preceding claims, comprising outputting information on haplotype frequency in the sample.

18. A method according to any one of the preceding claims, comprising outputting information on the haplotypes in the population from which the sample was taken.

19. A method according to any one of the preceding claims, wherein the haplotype information is obtained using the approximation that the conditional distribution of an undetermined haplotype h, given a set of assumed-correct haplotypes H, is

$$\sum_{\alpha \in E} \sum_{s=0}^{\infty} \frac{r_\alpha}{r} \left(\frac{\theta}{r+\theta} \right)^s \frac{r}{r+\theta} \left(P^s \right)_{\alpha h}$$

where r_α is the number of haplotypes of type α in the set H, r is the total number of haplotypes in H, θ is a scaled mutation rate, and P a mutation matrix.

20. A method according to any one of the preceding claims, comprising:

starting with an initial haplotype reconstruction $H^{(0)}$;

iteratively for $t = 0,1,2,\ldots$, obtain $H^{(t+1)}$ from $H^{(t)}$ using the following three steps:

1. choose an individual i from all ambiguous individuals;

2. Sample $H_i^{(t+1)}$ from $\Pr(H_i | G, H_{-i}^{(t)})$, where H_{-i} is the set of haplotypes excluding individual i;

3. Set $H_j^{(t+1)} = H_j^{(t)}$ for $j = 1,\ldots,n, j \neq i$.

21. A method according to any one of the preceding claims, comprising:

starting with an initial guess H for the haplotype reconstructions of all individuals, make a list consisting of the haplotypes $h=(h_1,...,h_m)$ present in H, together with counts $r=(r_1,...,r_m)$ of how many times each haplotype appears; iterate the following four steps:

1. Choose an individual i, and remove its two current haplotypes from the list (h,r) (so the list now contains the haplotypes in H_{-i}); let k be the number of loci at which i is heterozygous;

2. Calculate a vector $p=(p_1,...,p_m)$ as follows: for $j=1,...,m$ check whether the genotype G_i could be made up of the haplotype h_j plus a complementary haplotype, h' say; if not, set $p_j=0$, but if so search for h' in the list $(h_1,...,h_m)$; if h' is in the list, $h'=h_k$ say, then set $p_j=(r_j+\theta/M)(r_k+\theta/M)-(\theta/M)^2$, otherwise set $p_j=r_j(\theta/M)$;

3. With probability $2^k(\theta/M)^2\Big/\left(\sum_j p_j + 2^k(\theta/M)^2\right)$ reconstruct the haplotype for individual i completely at random (i.e. by randomly choosing the phase at each heterozygous locus); otherwise reconstruct the haplotype for individual i as h_j plus the corresponding complementary haplotype, with probability $p_j\Big/\sum_{j'} p_{j'}$;

4. Add the reconstructed haplotype for individual i to the list (h,r).

22. A method according to any one of the preceding claims, comprising:

starting with an initial haplotype reconstruction $H^{(0)}$; iteratively, for $t=0,1,2,...$, obtain $H^{(t+1)}$ from $H^{(t)}$ using the following three steps:

1. Choose an individual i from all ambiguous individuals;

2. Select a subset S of ambiguous loci or sites in individual i to update; let $H(S)$ denote the haplotype information for individual i at the loci or site in S, and $H(-S)$ denote the complement of $H(S)$, including haplotype information on all other individuals (so $H(S) \cup H(-S)=H$); sample $H^{(t+1)}(S)$ from $\Pr(H(S)\,|\,G, H^{(t)}(-S))$;

3. Set $H^{(t+1)}(-S)=H^{(t)}(-S)$.

23. A method according to anyone of the preceding claims, wherein the method is modified to deal with unknown or missing genotype data.

24. A method according to claims 23, wherein steps 1 to 3 are replaced by the following:

Step 4: Choose an individual i uniformly randomly from all individuals;

let $\overline{H}_{(i,c)}$ denote the unknown haplotype data in individual i, chromosome c;

Sample $\overline{H}_{(i,0)}^{(t+1)}$ from $\Pr\left(\overline{H}_{(i,0)} \middle| H^{(t)}, G^{(t)}, \overline{H}_{-i}^{(t)}\right)$;

sample $\overline{H}_{(i,1)}^{(t+1)}$ from $\Pr\left(\overline{H}_{(i,1)} \middle| H^{(t)}, G^{(t)}, \overline{H}_{-i}^{(t)}, \overline{H}_{(i,0)}^{(t+1)}\right)$;

set $\overline{H}_{-i}^{(t+1)}$ to $\overline{H}_{-i}^{(t)}$; and

update $\overline{G}^{(t+1)}$ to be consistent with $\overline{H}^{(t+1)}$,

where $G = (G_1, \ldots, G_n)$ denotes the *known* genotype data, $\overline{G} = (\overline{G}_1, \ldots, \overline{G}_n)$ denotes the *unknown* genotype data, $H = (H_1, \ldots, H_n)$ denotes the haplotype data corresponding to the known genotype data, and $\overline{H} = (\overline{H}_1, \ldots, \overline{H}_n)$ denotes the haplotype data corresponding to the unknown genotype data, and where the full genotype data $(G, \overline{G})$ are used in steps 1 and 2 of the algorithm.

25. A method according to claim 24, wherein at each iteration either steps 1 to 3 are performed or they are replaced by step 4.

26. A method according to claim 25, wherein successive iterations alternate between performing steps 1 to 3, and performing step 4.

27. A method according to any one of the preceding claims, wherein the conditional distribution is approximated by:

$$\Pr(H_i | G, H_{-i}) = \sum_{(n_1,c_1)} \sum_{(n_2,c_2)} \sum_{t_1=1}^{Q} \sum_{t_2=1}^{Q} \left[\Pr(n_1, c_1, n_2, c_2, t_1, t_2 | G, H_{-i}) \prod_{r=1}^{R} \Pr\left((h_{i0r}, h_{i1r}) | G, H_{-i}, n_1, c_1, n_2, c_2, t_1, t_2 \right) \right]$$

$$= \sum_{(n_1,c_1)} \sum_{(n_2,c_2)} \sum_{t_1} \sum_{t_2} \frac{1}{(2N-2)} \frac{1}{(2N-1)} W_{t_1} W_{t_2}$$

$$\prod_{r=1}^{R} F\left(h_{n_0,c_0,r}, h_{i,0,r}; \theta, T_1, 2N-2\right) F\left(h_{n_1,c_1,r}, h_{i,1,r}; \theta, T_1, 2N-1\right)$$

where R is the total number of loci/sites, $h_{n,c,r}$ is the allele carried by individual n at locus r on haplotype c ($c = 0, 1$), the sum over indices (n_1,c_1) is over all haplotypes in H_{-i}, the sum over indices (n_0,c_0) is over all haplotypes in $H_{-i} \cup h_{i,0}$,

$$F(i, j; \theta, t, n) = \sum_{m=0}^{\infty} \frac{(\theta t / n)^m}{m!} \exp(-\theta t / n)\left(P^m\right)_{ij},$$

there are Q quadrature points $(T_1,\ldots,T_Q)$ and associated quadrature weights $(W_1,\ldots,W_Q)$ and n_1,n_2,c_1,c_2,t_1,t_2 are latent variables.

28. A method according to claim 27, wherein $Q = 4$, and the sum F is approximated by summing over a predetermined number of terms.

29. A method according to claim 27 or 28, further comprising the step of: initially computing values for the function F, and powers of F, and storing them in a look-up table.

30. A method according to any one of claims 20 to 29, wherein at step 1 the individual i is chosen at random at each iteration.

31. A method according to any one of the preceding claims, further comprising the steps of: initially storing in an array the number of differences between every pair of haplotypes at biallelic loci; and at the end of each iteration, only updating the elements of this array that have changed during the iteration.

32. A method according to any one of the preceding claims, wherein the genotype information is sequence data of at least 5 segregating sites, preferably at least 30 segregating sites, more preferably at least 60 segregating sites.

33. A method according to any one of claims 1 to 31, wherein the genotype information is from linked microsatellite loci, preferably at least 5 linked microsatellite loci, more preferably at least 10 linked microsatellite loci.

34. A method according to any one of claims 1 to 31, wherein the genotype information is SNP data, preferably 10-30 SNPs.

35. A system for determining haplotype information from genotype information on individuals in a sample, comprising:

an interface for receiving genotype information;

a module for executing a Markov chain Monte Carlo algorithm to derive information on the conditional distribution of haplotypes, based on the genotype information;

a module for estimating haplotype information using the derived information on the conditional distribution; and

an interface for outputting said haplotype information.

36. A system according to claim 35, arranged to perform the method of any one of claims 1 to 34.

37. A computer program which is capable, when executed by a computer processor, of causing the computer processor to perform a method according to any one of claims 1 to 34.

38. A computer-readable storage medium having recorded thereon a computer program according to claim 37.

Fig.1.

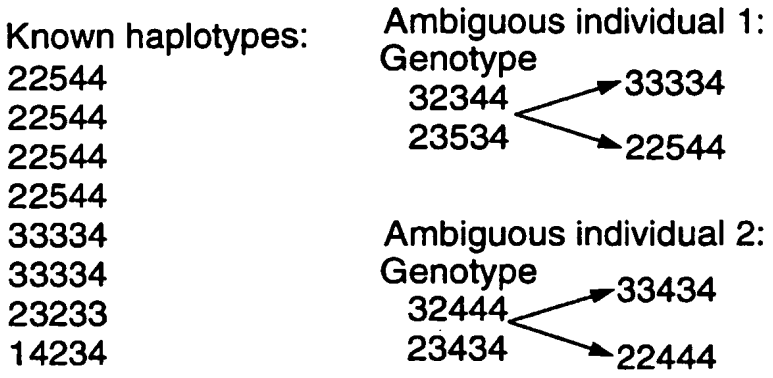
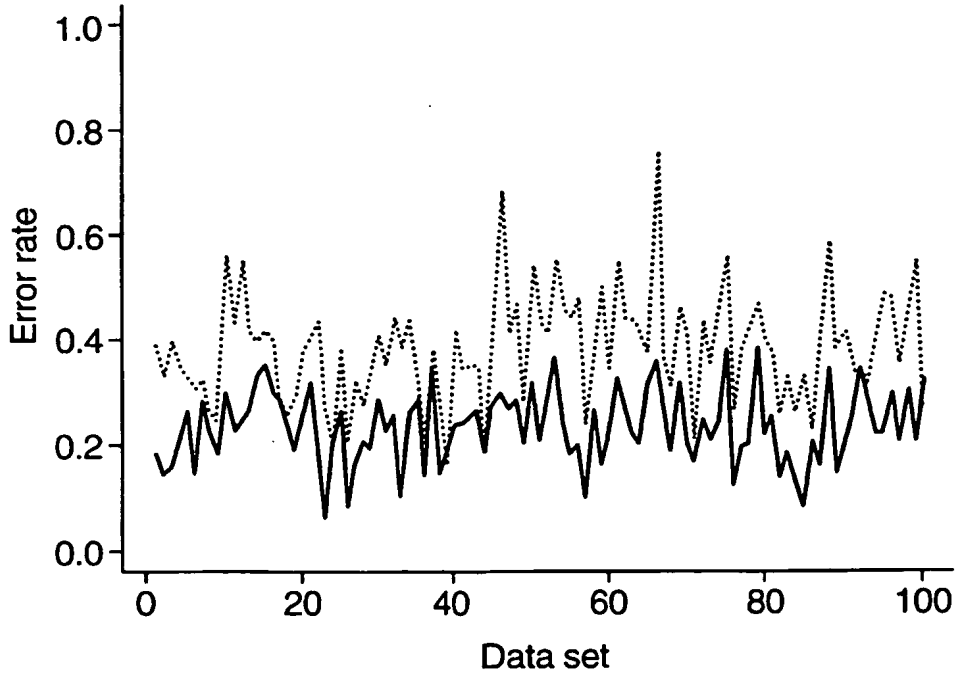

Known haplotypes:
22544
22544
22544
22544
33334
33334
23233
14234
Ambiguous individual 1:
Genotype
32344
23534
33334
22544
Ambiguous individual 2:
Genotype
32444
23434
33434
22444

Fig.4.


Error rate
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
Data set

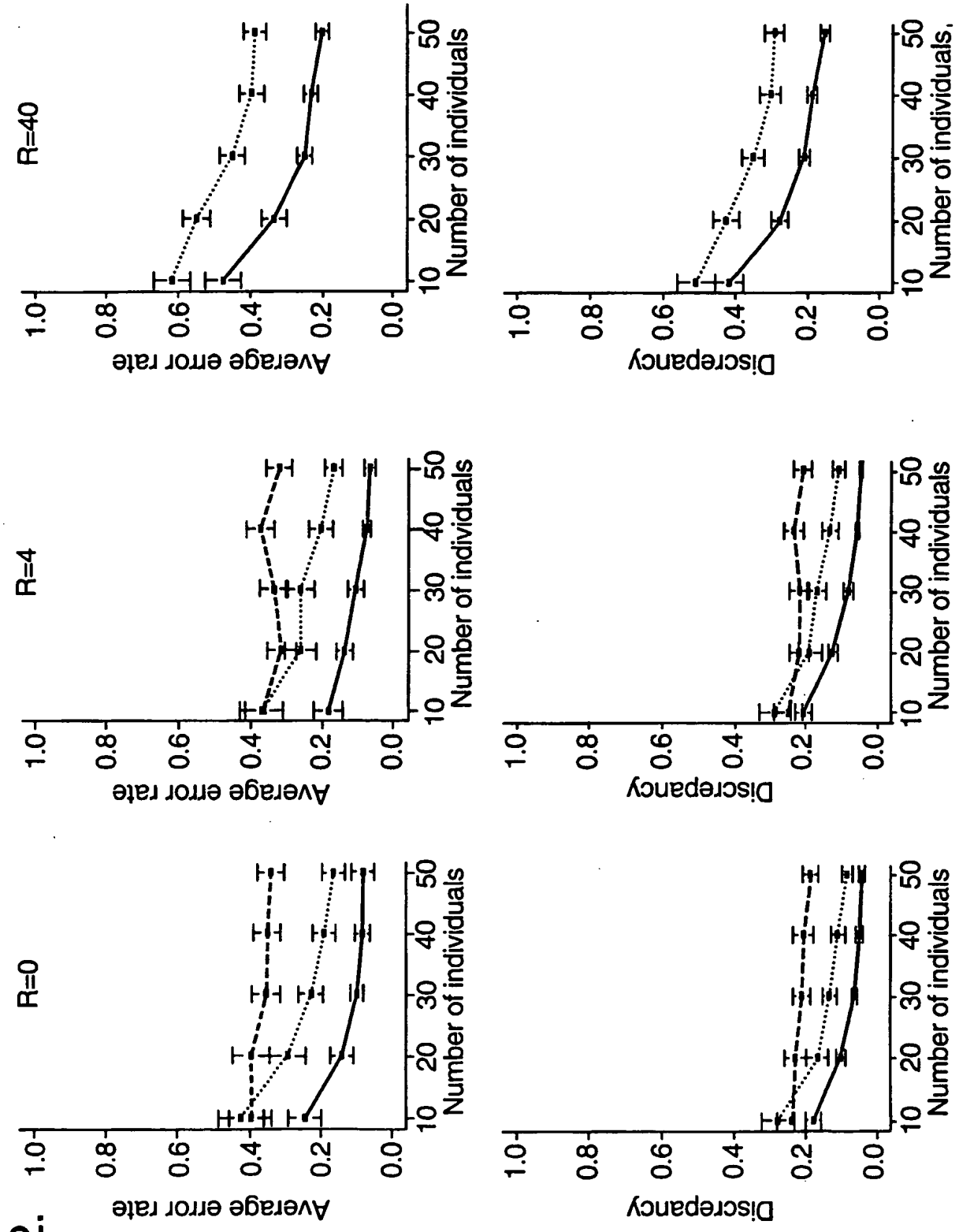

R=0
R=4
R=40
Average error rate
Discrepancy
Number of individuals
1.0
0.8
0.6
0.4
0.2
0.0
10
20
30
40
50

Fig.2.

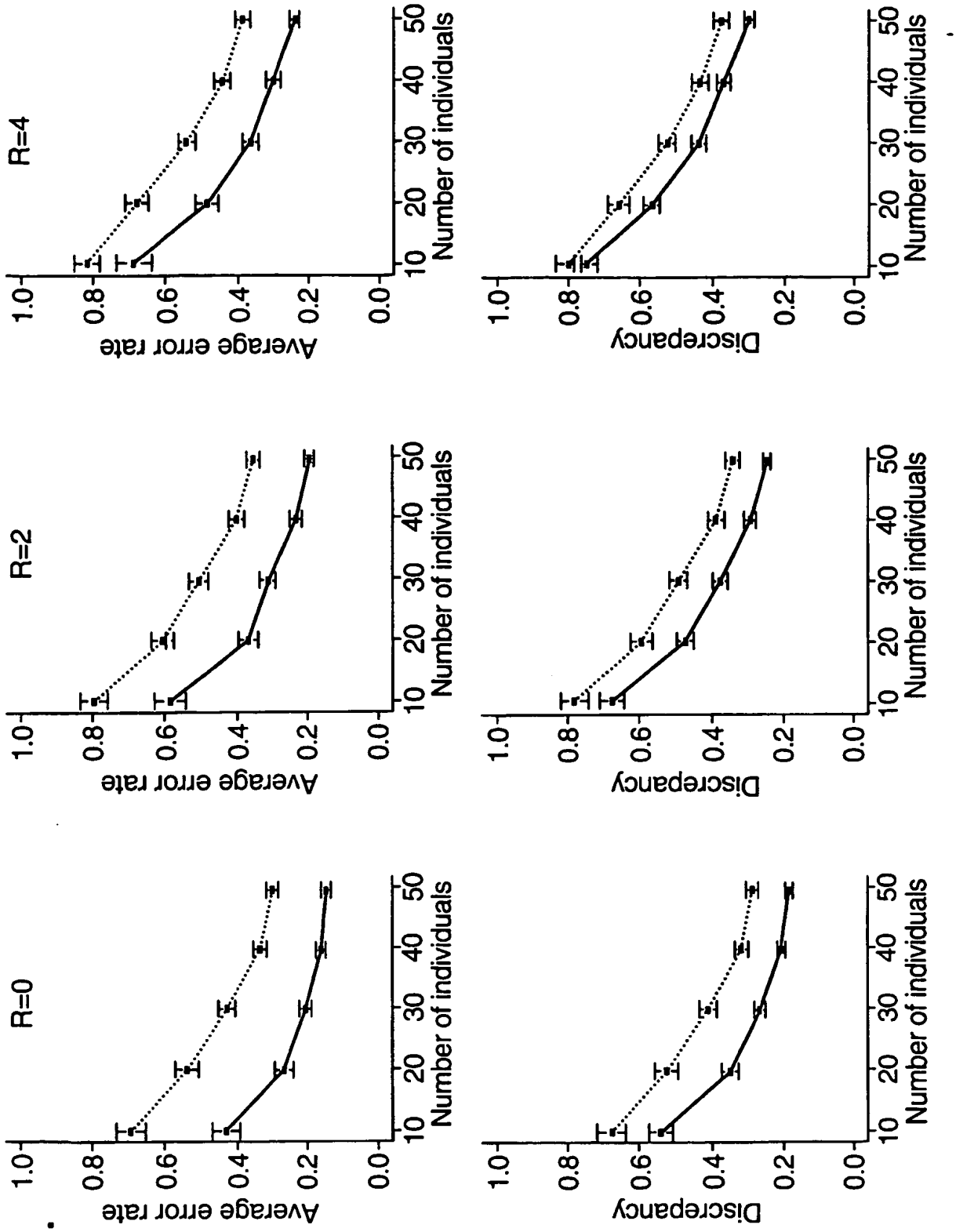

R=0
R=2
R=4
Average error rate
Discrepancy
Number of individuals
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50

Fig.3.